Corresponding lexical domains: A new resource for onomasiological bilingual dictionaries

Trklja, Aleksandar

DOI:
10.1093/ijl/ecw019

Document Version
Peer reviewed version

Citation for published version (Harvard):
Trklja, A 2016, 'Corresponding lexical domains: A new resource for onomasiological bilingual dictionaries', *International Journal of Lexicography*, vol. 29, no. 3. https://doi.org/10.1093/ijl/ecw019

Corresponding lexical domains: A new resource for onomasiological bilingual dictionaries

Abstract

The purpose of the current study is to develop a model for analysing relations between lexical items across languages in terms of lexical domains. The model combines a corpus-informed distributional approach with the language-in-use theory of meaning to identify sets of semantically similar linguistic items across languages. It also applies the principle of differentiation to establish differences between individual items. The model overcomes the limitations of previous methods which rely heavily on introspection, and focus on single words. It will be proposed that the results of the study can be used as resources in the compilation of a new type of onomasiological bilingual dictionary. Such a dictionary would provide users with direct access to multi-word units across languages, and help them distinguish and choose between available options.

1. Introduction

According to Goddard and Thieberger (1997), the semasiological dictionary presently dominates both the market and the field of lexicography. Hüllen (1999) demonstrates that in previous centuries the onomasiological bilingual dictionary had a more important role in lexicography than nowadays. For example, out of 1858 dictionaries published between 1467 and 1600 for various European languages, 475 were onomasiological dictionaries (Claes, 1977). In fact, some of the earliest bilingual dictionaries in Europe consisted of translations of Latin terms into other languages, and the terms from both languages were arranged thematically into lists. For example, one of the most important dictionaries in the history of English lexicography, *Ælfric's Glossary*, was an onomasiological dictionary for the language-pair Latin/Old English. Following Hartmann and James (1998: 101) we can define the onomasiological dictionary as "[a] type of REFERENCE WORK which presents words or phrases as expressions of semantically linked CONCEPTS, which may be meanings, ideas, notions, word families and similar relationships". As these and other authors point out (e.g. McArthur, 1998; Sterkenburg, 2003), the main difference between onomasiological and semasiological dictionaries is that in the former we move from concept to word, and in the latter from word to the explanation of a concept.

Although various studies[1] demonstrate that bilingual dictionaries are in at least some respects more useful than monolingual dictionaries, they have not undergone the tremendous development seen in monolingual dictionaries since the advent of text corpora. This does not mean that corpora have not been used in bilingual dictionaries[2], but they still lack the information about the context in which words occur. The entries in these dictionaries are typically single-word based, and collocations included in entries are mainly chosen arbitrarily. The following statement discussing German-English bilingual dictionaries sums up neatly the problem of bilingual dictionaries in general:

> "Conventional bilingual dictionaries are often little help here, as they frequently give a fairly undifferentiated list of possible German equivalents for a particular English word without providing much detail on how those German equivalents are actually used or the types of context where one might be preferred to another." (Durrell, 2000: x)

Dictionary-makers, therefore, "shift the burden of choice to the user of the dictionary" (Martin, 1967: 56). It is questionable whether significant advances are possible in this area as long as bilingual dictionaries are based on what Snell-Hornby (1987) calls *the approximation principle*. She argues that, instead, we should adopt the *principle of differentiation* and combine it with the theory of semantic fields. Dictionaries based on this approach would:

> "aim at pinpointing the focal components of the lexeme concerned and at situating it both paradigmatically (or intralingually) and contrastively (or interlingually), i.e. both against other items in the semantic fields concerned and in contrast to similar items in the target language." (Snell-Hornby, 1990: 222)

In another paper the same author suggests that translation dictionaries should be "contrastive dictionaries of synonyms, whereby the alphabetical system gives way to arrangement in semantic fields" (Snell-Hornby, 1984: 278). Such dictionaries would provide an explanation of words both in terms of the synonymy and equivalence relation.

[1] See, for example, Atkins and Knowles (1990), Baxter (1980), Nord (2002) and Tomaszczyk (1979). A comprehensive comparison of semasiological and onomasiological dictionaries is provided in Siepmann (2006). For a cognitive onomasiological approach to lexical semantics see Grondelaers and Geeraerts (2003).
[2] For more details on how corpora are used in bilingual lexicography see Atkins (1994).

What is required is a new approach to meaning. Zgusta (Wierzbicka, 1987: 1–2) argues that "if the treatment of meaning in dictionaries is to be radically improved, preparatory work has to be done by linguists". In the present paper, one example of such 'preparatory work' will be presented. I will argue that this goal can be achieved by a model that combines corpus methods with the distributional theory of meaning. The focus will be on the development and testing of a new theoretical and methodological framework, rather than on the design of dictionaries. The model of analysis comprises three major steps. First, semantic domains[3] will be identified through an investigation of the occurrence of lexical items in a parallel corpus. Second, lexical items from the same domain will be semantically differentiated by means of a local grammar analysis. Finally, relations between lexical items that belong to the same semantic group will be quantified and calculated, selection between them will be represented using the concepts from decision theory, and made available to dictionary-makers.

2. Previous studies[4]

Under Francis Bacon's influence, classifications of words into semantic groups have relied for centuries on the assumption that words stand for concepts or mental representations. Accordingly, scholars have regarded classifications of these concepts as representations of our knowledge of the world. At the core of such classifications lies a limited number of 'basic concepts' which are considered universal for all languages and to which the whole of our knowledge can be reduced (McArthur, 1986). It was believed that the dictionaries based on this view provided "their users with a view of the world" (Hüllen, 2009: 109). They were similar to encyclopaediae, to which they were historically related (Hüllen, 1999: 65). A taxonomy that Bacon proposed has had a tremendous influence on the history of lexicography. Some of his categories can be found in as diverse a range of sources as Comenius' *Orbis* (published in 1658), Roget's (1852) *Thesaurus*, Sanders' (1873) *Deutscher Sprachschatz*, McArthur's (1981) *Longman Lexicon of Contemporary English, Longman Language Activator* (Summers, 1993), *Cambridge Word Routes Anglais-Français: Lexique thématique de l'anglais courant* (McCarthy, 1994), *Diccionario temático del inglés*

[3] The meaning of the term *domain* is explained in Section 3.

[4] Needless to say, onomasiological studies and classifications of words into semantic groups have been carried out in other disciplines as well (e.g. CLIR, Terminology and Knowledge Engineering, Natural Language Processing). The present review, however, focuses only on several linguistic approaches relevant to the model proposed in the paper and due to space limitations it cannot systematically elaborate on approaches from other disciplines.

contemporaneo (Walter, 1995), *Cambridge Word Routes Anglika-Ellinika* (McCarthy, 1996), Dornseiff's (2004) *Wortschatz nach Sachgruppen* or *Historical Thesaurus of the Oxford English Dictionary* (Kay et al., 2009).

A more linguistically-informed method for the classification of words into semantic sets was introduced by the German linguist Jost Trier (1931) in his theory of *lexical-* or *semantic field* (in German *Wortfeld*). What is common to different implementations and formulations of this method that have evolved ever since is the view that the lexicon is an organized entity and that "the vocabulary of a language is structured, just as the grammar and phonology are structured" (Lehrer, 1974: 15). I will discuss here three approaches to lexical fields which are relevant for the present research[5].

When the theory of lexical fields was at its height (from the 1960s to the 1980s) it was dominated by the view that words consist of semantic components. The following quotation summarises this view.

> "A minimal definition of the meaning of an item will be a statement of the semantic components necessary and sufficient to distinguish the meaning paradigmatically from the meanings of all other items in the language." (Bendix, 1971: 393)

Differences between synonymous words that belong to the same semantic field were established through a componential analysis. Componential analysis seems at first sight suitable for the implementation of the principle of differentiation mentioned above. The model, however, suffers from several shortcomings and has been the subject of thorough critiques (e.g. Dixon, 1971; Geeraerts, 2010; Lyons, 1995; Van Roey, 1990). I will address some of its weak points by discussing Lehrer's (1974) study of words from the semantic field *cooking*.

Lehrer starts with an analysis of semantic components that underlie the meaning of words from her semantic field (e.g. *the use of water, the use of oil, cooking time* and *the use of cooking utensils*). She suggests that semantic differences between these words become clear when the distribution of these components is compared. She demonstrates, for example, that the difference between the verbs *boil* and *fry* is that only the former is associated with the

[5] For more details on other methods and approaches such as Viberg's study of semantic fields (1983, 1993) or Sinclair's study of translation equivalents (1996a, 1996b) see Trklja (2013).

component *cooking with the use of water* and that only the latter is associated with the component *cooking with the use of oil*. By the same token, *fry* differs from *sauté* because only the latter contains the component *the use of cooking liquid*. In a similar fashion, she suggests, we can also establish cross-linguistic relations between words. For example, the German verb *kochen* corresponds both to *cook* and *boil* in English because it contains both the component *the general process of cooking* and *the process of cooking with water*. Similarly, *braten* is an equivalent to *fry* and *broil* because it is associated with the components *the use of oil* and *no use of oil*.

The first problem with componential analysis has to do with the ontological nature of semantic components. As Gordon (2003: 2219) correctly notes, "[t]here is nothing to suggest the existence of any objective or universally applicable means of establishing parameters for a componential analysis". It is no accident that componential studies are usually limited to words that refer to concrete objects, basic activities and stative adjectives (Snell-Hornby, 1990: 211). It is relatively easy to identify semantic borders for such words "[b]ut many other vocabulary terms refer to 'things' which have features that are not neatly distinguishable, so that their meanings have 'fuzzy edges', i.e. contrast only vaguely and cannot be adequately described in terms of components" (Van Roey, 1990: 30). In addition, it is not clear how many components are required to describe the meanings of words that occur in the same field. Unlike in phonetics, where the componential analysis was initially developed, there is no way in semantics to establish in advance a limited number of universally applicable features. Often a very large number of components must be listed which makes the approach uneconomic (Dixon, 1971: 441). Lehrer (1974: 201) admits that "the problem of determining the inventory of the lexical items in a field remains". Finally, componential analysis focuses on isolated words and does not provide information about the contexts in which they are used. In some cases this can paint a wrong picture. There are numerous componential studies that demonstrate the existence of lexical gaps across languages. But, as Durrell (1981) rightly points out, the ostensible lexical incongruence and lexical gaps discussed by semantic field scholars are in reality products of the componential method itself, and not of differences between languages. Following the componential approach one would be reluctant to conclude that there is no translation equivalent of the English verb *simmer* in German. But this is of course not true, because such an equivalent does exist, except that it is expressed in the form of a two-word lexical item *langsam kochen*. The same meaning is simply lexicalised in different ways in the two languages, which is not an unknown phenomenon in linguistics. According to Lyons (1977: 262), "[i]n many cases, one language will use a syntagm where

another language employs a single lexeme with roughly the same meaning". Lehrer notes that the approach might not be well equipped to deal with complex words, phrases and idioms and concludes that "[t]his may turn out to be a fundamental mistake" (Lehrer, 1974: 201). For all these reasons, the traditional componential method cannot be accepted for the purposes of the present study.

A more recent approach to semantic fields proposed by Dyvik (1998, 2004, 2005) remedies some of the limitations of the traditional componential method. The author starts from the assumption that the meaning of words becomes visible in translation. The fact that perfect translations are impossible and that the target language is always "like a Procrustean bed for the source language" (Dyvik, 2005: 7) should not be regarded as a drawback, according to him. On the contrary, it is the difference between languages that can provide interesting insights about the semantics of words. In addition, Dyvik (2004: 1) argues that "semantically closely related words ought to have strongly overlapping sets of translations, and words with wide meanings ought to have a higher number of translations than words with narrow meanings".

Unlike in the conventional componential model where relations between words are established mainly through intuition or in some case by means of dictionaries, Dyvik proposes an approach that relies on the occurrence of lexical items in a parallel corpus. First, translation correspondences of a word from a source language (SL) are identified in a target language (TL) in a parallel corpus. After that, these words are translated back into SL and individual senses are identified through the observation of the distribution of words. Individual words are regarded as t-images (translation images) that might have overlapping senses. The higher the number of t-images in which a sense occurs, the wider the meaning words will have. Semantic differences and similarities between these words are established through the comparison of the distribution of senses. I will illustrate the method with several words from the semantic field *meal* dealt with by the author in Dyvik (2005). The senses associated with the Norwegian nouns *aftensmat*, *kveldsmat* and *måltid* are displayed in the following table:

Senses/Words	*aftensmat1*	*kveldsmat1*	*måltid1*
[mat1|supper2]	√	√	√
[lunsj1|meal1]	√		
[kveldsmat1|meal1]	√	√	
[aftensmat1]	√		
[måltid1|dish3]			√

Table1: *Semantic field 'meal'*

Words are displayed here in the first row and senses in the first column, both as Norwegian and English terms. Numbers denote what Dyvik refers to as 'sense partitions'. Thus, [mat1|supper2] indicates that the first sense of the word *mat* corresponds to the second sense of the word *supper*. It can be observed in the table that the first word (*aftensmat)* is associated with four different senses and other two words (*kveldsmat and måltid)* with two. The table also shows that both senses of the word *kveldsmat* occur also with *aftensmat*, and that only one sense associated with *måltid* occurs with *aftensmat*. It follows that the three words are semantically similar but not identical; the first word has the most general meaning, the second word shares all senses with the first word, and the third word is used in one sense like the other two words and in one sense idiosyncratically.

The most important advantage of this approach in relation to the traditional componential method is that identification of relations between words across languages does not rely on intuition but on the observation of occurrences of words in a parallel corpus. Unfortunately, this approach also focuses only on single words and ignores the context in which words occur. For this reason, it fails to differentiate appropriately between synonymous terms that belong to the same semantic field. For example, according to the above description, the words *aftensmat*, *kveldsmat* and *måltid* are regarded as synonyms when they are used in the sense coded as [mat1|supper2]. However, we do not know whether they are interchangeable in all textual contexts and whether they can co-occur in this sense with identical collocations. The same problem arises in a cross-linguistic description. Dyvik (2004), for example, describes one sense of the English adjective *sweet* in terms of the Norwegian item *frisk.* Nevertheless, the meaning of *frisk* may vary depending on collocations and we do not know when it exactly corresponds to *sweet*. For example, *frisk* collocates with *luft* and according to the multilingual OpenSubtitles corpus its English equivalent for this collocation is *fresh air* and not *sweet air*. In addition, the concordance lines from the BNC and ukWaC show that it is *fresh air* which is an idiomatic expression in English and not

sweet air. No such information is available in Dyvik's description. It seems that the main methodological procedure relies on the assumption that the meaning of a word can be defined in terms of another word. But, if we do not know how exactly the latter word is used then the whole enterprise becomes highly problematic. Words outside the textual context are polysemic and it follows that the meaning of one word is defined in terms of another semantically ambiguous word. This returns us to the age-old issue of circularity of definitions and to the aforementioned problem of shifting of the burden of selection to dictionary users.

A few words should be said also about an approach proposed by Siepmann (2005). The aim of his study, like ours, is to create a methodology which could be applied in the compilation of bilingual onomasiological dictionaries. In addition, his focus is on multi-word units rather than on single words. The following question is central for Siepmann: "what are the meaning units that native speakers use, and which of these have to be mastered to be able to perform at a near-native (or lower) proficiency level?" (Siepmann, 2005: 4). The principal units of analysis here are topics that can be divided into sub-areas. One such topic is, for example, *motoring* and one of the sub-areas that it encompasses is *parking*. The most important finding of this study is that the choice of equivalents across languages depends on the context in which lexical units occur.

Unfortunately, the notion of topic is ambiguously explained in terms of the concept of situation-type, which itself is only vaguely defined. Additionally, the identification of topics and equivalence relations is based on introspection. Similarly, in the extraction of collocates Siepmann rejects the statistical method and suggests that it is the analyser's feeling for language that should govern our decision as to what constitutes a near-native expression. As a result we have a highly subjective definition of topics, equivalents and collocations.

The theory of frame semantics and FrameNet has been proposed as an alternative to the traditional theory of semantic field frame semantics. This theory goes beyond the mere terminological descriptions of semantically similar words and accounts for syntagmatic relations between components that characterise the meaning of words (e.g. Fillmore and Atkins, 1992). The basic units of analysis are frames and features. For example, two features which characterise the frame for the noun *car* are ENGINE and DRIVER. Their syntagmatic and semantic relation is defined as: DRIVER controls the ENGINE as in the following invented example where *she* is DRIVER and *my old car* is ENGINE.

She drove my old car.

A serious limitation of frame semantics lies in the fact that semantic labels are based on *a priori* established categories. The problem is that the criteria that underlie these categories are not stable. In cognitive psychology it is held that frames "are continually updated and modified due to ongoing human experience" (Evans and Green, 2006: 223). As Hanks (2004: 6) notes, frame semantics "requires the researchers to think up all possible members of a Frame a priori, [which] means that important senses of words that have been partly analysed are missing and may continue to be missing for years to come." These omissions are discussed in Hanks (2004) and in Hanks and Pustejovsky (2005). One example discussed is the verb *toast* which is described only in terms of the Apply_Heat frame and it follows that frame semantics recognises only its sense associated with cooking and ignores the sense associated with the field of celebration. The reliance on pre-established categories, therefore, bears the risk of neglecting certain aspects of word meanings. The theory of frame semantics was initially developed only for English, but was subsequently also applied to other languages (e.g. Boas, 2002; 2005; Braasch, 1994; Heid, 1996). Frames used in these studies are adopted from English. However, this strategy overlooks the possibility that lexical units in other languages might evoke frames that do not exist in English. According to Hanks (2004: 6) "What is needed is a principled fix – a decision to proceed from evidence not frames. This is ruled out by FrameNet for principled reasons: the unit of analysis for FrameNet is the frame, not the word".

WordNet is a semantically organised lexical database of English (Fellbaum 1998). Words are here classified by part-of-speech categories and then grouped into synonym sets (synsets). Each word is represented by its senses, synset, definition, examples, sentence frames and other types of semantic relations (hyponyms, hyperonyms, antonyms, etc.). Grouping of words into synsets is based on arbitrary decisions and "[f]rom the WordNet literature available, it is often difficult to determine the bases on which design decisions in WordNet are made" (Murphy, 2003: 111). The database consists mainly of individual words but it contains also some collocations which seem to be selected arbitrary. Due to the focus on individual words WordNet is not very helpful for meaning disambiguation of multiword expressions. Teubert (2004) demonstrates this in his analysis of the entries of *friendly* and *fire* according to which these words have four and eight senses, respectively. It follows that the expression *friendly fire* is associated with thirty-two possible senses. In reality, when these two words are analysed as a collocation and not as a "contingent co-occurrence of two single words" (Teubert, 2004: 145) it becomes clear that they form one semantic unit.

Unlike traditional semantic fields WordNet displays also syntagmatic relations in the form of 'Sentence frames'. However, these frames provide only a general description and do not reflect the actual local contexts in which terms occur. For example, the verb *cause* in WordNet is associated with the sentence frame: *Somebody causes something*. From this, one can only conclude that the verb requires a subject and an object or complement, which is not very informative given the fact that many other verbs also occur in this general context.

Similar databases have been compiled for numerous other languages, one such being EuroWordNet, created for several European languages. Unlike in FrameNet, semantic categories (synsets) are here considered as language-specific, which means that they are not automatically adopted from English. Relations between synsets across languages are established via 'Inter-Lingual-Index' (ILI). Although in theory this link should be based on a neutral interlanguage, in reality ILI is mainly based on WordNet (Vossen, 2004). This database shows that words from different languages do not have exactly the same number of senses, meaning that ILI will not capture adequately the semantics of words from other languages if a particular sense does not exist in English. In addition, we saw above that WordNet does not treat multiword expressions adequately, and we may expect that EuroWordNet or other similar multilingual platforms will increase the number of senses even furtherwhen these are multiplied across languages. It should, finally, be pointed out that this model does not account for similarities and differences in uses of words from same synsets.

3. **Theory**

Following previous approaches, I will start from the assumption that lexical items can be classified into semantic groups. In addition, I will adopt Dyvik's model of studying correspondence relations across languages by means of parallel corpora. I will also take into account the findings following from Siepmann's analysis, that correspondence relations between lexical units in two languages change when the items are observed in different contexts. In order to address adequately the issues of textual contexts and principle of differentiation I will additionally introduce a few new theoretical and methodological notions.

In previous approaches, the notion of *meaning* was defined referentially either in terms of ideas placed in the mind or things that exist somewhere outside language. For Harris (2005), these naïve views are part of the same language myth, and none of the two assumptions can be proved to be true or false. In contrast, I will instead follow the language-

in-use theory of meaning introduced by Wittgenstein (1953) and further developed for the purposes of linguistic studies by Firth (1968) and Sinclair (1991). According to this view, the meaning of a word is defined in terms of its distribution. It is assumed that the meaning of a lexical item varies depending on the textual or situational context in which it occurs. Due to space limitations, in the present paper only the former type of context will be explored.

The process of identification of semantically similar multi-word items will rely on two assumptions. The first assumption, known as *distributional hypothesis*, claims that

> "if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words: difference of meaning correlates with difference of distribution" (Harris, 1970: 785).

Distribution of an element is here defined as all textual environments in which the element occurs (Harris, 1952; 1954). Harris (1952) further argues that lexical items that occur in the same context create an equivalence class. What qualifies items to belong to an equivalence class is that they are substitutable for each other. Historically, substitutability as an indicator of the equivalence relation is reminiscent of Leibniz's definition of identity relations: "Two things are the same if one can be substituted for the other without affecting the truth" (Lyons, 1977: 160).

The second assumption implies that "[t]here are likely to be parallels between the textual environment of a word in one language and a word that is used to translate it in another" (Sinclair, 1996b: 179). In other words, a close investigation of shared textual contexts in two languages should help to identify translation equivalents.

When the two assumptions are combined we come up with a new hypothesis for the identification of semantically similar items across languages: if two or more lexical items from language A occur in the same context and correspond to the same lexical items from language B, these items will be substitutable. The cross-linguistic semantic similarity of linguistic items is here defined in terms of interchangeability. It means that all lexical items that belong to the same semantic field will be regarded as belonging to the same substitution or equivalence class. This cross-linguistic distributional hypothesis stresses that equivalence sets should be established from both an intralingual and interlingual perspective. The original distributional hypothesis itself is not specific enough because lexical items that occur in the

same textual context are not automatically synonyms[6]. Such items might be antonyms, or they might belong to other types of semantic relations that have not been properly studied yet. For example, Church et al. (1994) show that words such as *pledge* and *contribute* may occur in the same context but are not synonyms. In order to make sure that such cases do not occur in our data the substitution sets need to be established first by means of an interlinguistic analysis and then confirmed through an intralinguistic analysis. No two items from L2 will correspond to the same item from L1 and simultaneously occur in the same context unless they are synonymous**.** Correspondences identified will create an equivalence class. These correspondences will, therefore, be defined as sets of lexical items with the same distribution. In this way we will come close to Zgusta's (2006: 236) lexicographical ideal that the task of lexicographers should be to find "real lexical units of the target language that, when inserted into the context, produce a smooth translation".

The principle of differentiation will be implemented in the present paper in terms of the local grammar approach (Gross, 1993). Local grammars are based on "a purely word-combinatorial investigation" (Harris, 1988: 40) and not on *a priori* created categories. Although the approach was introduced to study subject-matter specific domains such as immunology (Harris et al., 1989), technical manuals for aviation (Kittredge, 1982) or task-oriented dialogues (Grosz, 1982), I will argue that it can be extended to the study of all types of semantically restricted domains. General grammar categories are often too crude to describe distribution of lexical items. Waismann (1965) notes that general grammatical categories cannot explain why *north-east* does not occur in the contexts *x of the North Pole* and *x of the South Pole* such as in *north-east of the South Pole*. The general grammar description cannot say more than that *north-east* can be followed either by a noun or pronoun. The reason is that "our division of words into separate types probably follows principles that are too rough" (Waismann, 1965: 136). Waismann concludes that

> "it would be arbitrary to accept that it is a rule of grammar that 'north-east of' must be followed by a noun or pronoun in the accusative, yet to deny that it is a rule of grammar that these must themselves be designators of a place, an object or person at a place, or of an event occurring at a place." (Baker and Hacker, 2009: 63)

[6] One of the first systematic discussions of synonyms in lexical semantics is Cruse (1986).

Consequently, to provide a fine-grained description of the occurrence of words we must "dig deeper, pushing aside the outward division of words into noun, adjective, etc." (Waismann, 1965: 136) and create categories that describe local contexts in which words occur. The local grammar approach makes it possible to develop a precise description of the function of lexical items in a specific context. This is because the assigned category labels "are far more transparent than the highly general ones" (Hunston and Sinclair, 2000: 80). Not less importantly, with local grammars we can explain the function of multi-word units and go beyond the traditional single-words-centred parts-of-speech categories. A close comparison of local grammars of semantically similar words will show similarities and differences in their use.

To sum up, I will regard lexical items that occur in the same substitution set as semantically similar items. I will assume that the substitution sets identified by applying the cross-linguistic distributional hypothesis will correspond to each other. The items from these sets will be regarded as corresponding linguistic items. From the above observations it also follows that local grammars cannot only be applied to the study of subject-matter domains, but that it can define the very existence of semantic domains. Against this background, I will use the term *domain* to refer to sets of lexical items that share semantics and local grammars. The notion of *domains* is also more suitable than the traditional terms *lexical fields* or *semantic fields* because the latter have been associated with the referential theory of meaning.

I will use the notion of *correspondence* rather than the notion of *equivalence* because unlike the latter it does not presuppose that the items in two languages are semantically identical. Similarly, a technical term for the substitution sets of translation correspondences will be *corresponding lexical domains* (abbreviated as CLD). Finally, following Sinclair (2004: 132) I will talk about *lexical items* and not *words* because the former term is more adequate as a technical concept. I will also use the terms *lexical units*, *linguistic items* and *linguistic units* in the same sense.

4. **Analysis procedure, data and tools**

The model of analysis proposed here is not restricted to specific language pairs but because of the limitation of space only the language-pair English/German will be considered in the present paper.

The analysis involves three steps. First, corresponding lexical domains will be generated through an analysis of the occurrence of lexical items in a parallel corpus. In the second step, a local grammar analysis of lexical items from these domains will be carried out to describe the textual context in which they occur, and to identify similarities and differences between them. A reference corpus will be used in this step. The identification of corresponding lexical items and their collocates is carried out automatically by means of the corpus tools described below, but the analysis and annotation of relevant local grammars was conducted manually. Finally, the results following from the previous two analyses will be interpreted in terms of decision theory and thus made operable for lexicographical purposes.

The Europarl corpus (Version 6) which is a collection of discussions extracted from the proceedings of the European Parliament (Koehn 2005) will be used as a parallel corpus. The main reasons why this corpus is selected are its size and free availability. Translation correspondences are often multi-word units and other available parallel corpora such as INTERSECT (Salkie, 1995) are not big enough for the purposes of the present study. For example, *give rise to problems*, which is one of the terms that will be investigated below, occurs only twice in the INTERSECT corpus. A disadvantage of the Europarl corpus is that the original language in which texts are written is unknown, and the corpus is not representative of English and German in general because it is biased towards specific EU jargon. Both problems are related to the issue of typicality of expressions, and their potentially negative effect is controlled through the intralingual analysis of lexical items in a reference corpus.

The ukWaC corpus, which contains 1.9 billion tokens, will serve as a reference corpus for English. Although one may question its representativeness because it is a web-based corpus, Ferraresi et al. (2008) demonstrate that in terms of quality it matches the British National Corpus (BNC). Similarly, the deWaC, which is a German cognate of the ukWaC and consists of 1.7 billion tokens, will be used as a reference corpus for German. Both corpora have the merit of being very large. This is an important point because smaller corpora do not provide enough data to study in detail the local grammars of less frequent items.

The corresponding lexical items in the parallel corpus will be identified by means of ParaConc (Barlow, 2008). Texts from the parallel corpus are aligned at the sentence level and ParaConc identifies and displays all concordance lines in which corresponding items from two languages occur. The collocation analyses of lexical items will be carried out with the help of the RCQP package (Desgraupes and Loiseau, 2012) and Shell-based scripts. The

package combines various statistical packages with the tools available in the IMS Corpus Workbench (Evert and Hardie, 2011) and can be added to the programming language R. A function that was employed in the present study produces lists of shared collocates for two lexical items and shows the values of their collocation strength. This function is similar to the SketchEngine tool called Sketchdiff (Kilgarriff and Kosem, 2012), but its advantage is that it is not restricted to the comparison of one-word units.

There are a few points worth noting about the conventions used in this paper. All lemmata will be placed within angled brackets and all word forms will be italicised (e.g. *costs*, *difficulties*). A lemma can take the form either of a single word or a multi-word expression (e.g. <problem> and <give rise to problem>). The names of lexical domains will be written in capital letters and put between curly brackets. In choosing the names of domains I follow Apresjan's (2000: 217) suggestion not to use artificial terms for metalanguage, but rather the words from the object language that are intuitively comprehensible. The domains will be named after the most frequent lexical items from a domain and represented with capital letters, fonts at the 12-point size and curly brackets (e.g. {CAUSE PROBLEM}). The local grammar categories will be coded with capital letters, 10 point size fonts and square brackets. Alternative elements will be represented by a vertical bar (e.g. [PROCESS|DECISION] <cause> [INTENSIFIER|QUANTITY] <problem|difficulty>).

5. Analysis

5.1 Identification of translation correspondences

The study of relations between translation correspondences in the parallel corpus involves the following steps. First, a lemma from English is randomly selected and its common German corresponding items are identified in the Europarl corpus. After that, these German lexical items are used as search terms and the corresponding English expressions are found in the same corpus. This type of back-and-forth translation process is repeated as many times as necessary to identify the most frequent items from both languages. Items are considered for further investigation only if they a) correspond to at least two items from another language; b) occur at least five times in our corpus, and c) correspond to more than two percent of the occurrences of an item from another language.

Following the substitution principle introduced in Section 3, the lexical items that occur in the same contexts and correspond to the same items from another language are regarded as belonging to the same lexical domain.

I will illustrate the model through a study of lexical items from CLD {CAUSE PROBLEM}. At the beginning, the item <give rise to>[7] is randomly selected. It occurs 1368 times in the Europarl corpus and corresponds to 24 different German lexical items. These corresponding items are displayed in the first column of Table 2 below. The first row shows the most typical collocates of <give rise to> and the central part indicates in which contexts the lexical items from the two languages correspond to each other. For example, <give rise to> corresponds to <auftreten> only when it occurs with the nouns <problem> and <difficulty> and only when the German verb collocates with <Problem> and <Schwierigkeit>. Similarly, <stattfinden> corresponds to <give rise to> only when the former collocates with <Debatte> and the latter with <debate>.

These results indicate that the relevant semantic units in English consist of <give rise to> plus a small set of nouns, and in German of lexical items corresponding to these two elements. It can also be observed that the items from the two languages do not stand in a one-to-one relation. For example, <führen zu> and <Anlass geben> correspond to five collocations created with <give rise to>, and <entbrennen> and <wecken> to two such collocations. It means that the former German items have more uses in common with <give rise to> and occur in a higher number of similar contexts than the latter ones. Finally, given the fact that the lexical items from English and German do not occur in the same contexts and do not correspond to the same items, it can be concluded that they do not belong to the same CLD. For example, the verbs <führen zu> and <stiften> create one domain when they collocate with <Verwirrung> and correspond to <give rise to confusion>, whereas <führen zu> and <enstehen> constitute another domain when they collocate to *Kosten* and correspond to <give rise to> *costs*. Thus, in order to establish all domains in which <give rise to> occurs it would be necessary to investigate all German items in this manner and through the procedure described above in step two carry out a detailed analysis of all the items from two languages.

[7] As demonstrated in Trklja (2013), the model is applicable to the study of any lexical item regardless of its length, word class or polysemous nature.

	<give rise to problem>	<give rise to concern>	<give rise to fear>	<give rise to debate>	<give rise to confusion>	<give rise to difficulty>	<give rise to doubt>	<give rise to question>	<give rise to cost>
<zu Problem|Schwierigkeit|Sorge|Verwirrung|Kosten führen>	√	√			√	√			√
<Problem|Schwierigkeit auftreten>	√					√			
<Problem|Sorge|Debatte|Verwirrung auslösen>	√	√		√	√				
<Problem|Sorge|Debatte|Zweifel|Frage hervorrufen>	√	√		√			√	√	
<zu Sorge|Angst|Debatte|Zweifel|Frage Anlass geben>		√	√	√			√	√	
<Problem|Verwirrung|Schwierigkeit|Frage|Kosten entstehen>	√				√	√		√	√
<Problem|Schwierigkeit (mit sich) bringen>	√					√			
<Problem es gibt>	√								
<zu Debatte|Verwirrung kommen>				√	√				
<mit Problem|Schwierigkeite verbunden sein>	√					√			
<Debatte stattfinden>				√					
<Problem schaffen>	√								
< für Debatte|Verwirrung sorgen>				√	√				
<Debatte provozieren>				√					
<Debatte entbrennen>				√					
<Problem sich ergeben>	√								
<Problem|Kosten verursachen>	√								√
<Ursache sein>	√								
<Problem|Schwierigkeit|Frage aufwerfen>	√					√		√	
<Angst|Zweifel aufkommen>			√				√		
<Problem|Frage sich stellen>	√							√	
<Angst wecken>			√						
<Problem|Schwierigkeit bereiten>	√					√			
<Verwirrung stiffen>					√				

Table 2: <give rise to> and its German translation correspondences according to the Europarl corpus

Due to space restriction only one such domain will be examined below. The focus will be on the collocation <give rise to problem> because it has the largest number of translation correspondences in our corpus.

In the Europarl corpus, 14 German lexical items can be observed to correspond to <give rise to problem> (Table 3). A back-and-forth translation into English and German renders a list of additional correspondences which are summarized in Table 4.

German lexical items	Frequency in the Europarl corpus
<Problem bereiten>	3
<Schwierigkeit bereiten>	2
<Problem (mit sich) bringen>	4
<Schwierigkeit (mit sich) bringen>	2
<zu Problem führen>	16
<zu Schwierigkeit führen>	5
<Problem schaffen>	4
<Schwierigkeit schaffen>	1
<Problem verursachen>	8
<Problem aufwerfen>	6
<Schwierigkeit aufwerfen>	3
<Problem entstehen>	5
<Schwierigkeit entstehen>	2
<Problem ergeben sich>	4
<Ursache DET Problem sein>	4

Table 3: German translation correspondences for <give rise to problem>

English translation correspondences	
<cause of problem>	<pose problem>
<cause difficulty>	<present difficulty>
<cause problem>	<present problem>
<create difficulty>	<problem arise>
<create problem>	<raise difficulty>
<difficulty arise>	<raise problem>
<give rise to difficulty>	<result in difficulty>
<lead to difficulty>	<result in problem>
<lead to problem>	<there be difficulty>
<pose difficulty>	<there be problem>

Table 4: English lexical items corresponding to the German lexical units displayed in Table 3

A complete list of lexical items that occur at least twice in the corpus in both languages is provided in Appendix Tables I and II.

A further investigation of the relation between translation correspondences reveals some important facts. First, items are not involved in the same number of correspondence relations. Correspondence relations indicate here the number of lexical items from L1 to which an item from L2 correspond. Second, there are differences in probabilities with which items from L2 can be used as translation correspondences. There is a strong but not perfect

correlation (r > .80) between these two variables in the present data, and both variables should be taken into account. For example, a widely-used item from L2 may cover only two percent of the occurrence of an item from L1. On the other hand, it may also happen that an item from L1 covers 60% of the occurrence of an item from L1 but does not correspond to any other lexical item. I created a formula that considers both variables and calculates their values in comparable terms. The product of these values is called the *correspondence degree* and its value for the most frequent lexical items with the plural form of the nouns <problem> and <Problem> is displayed in Table 5 bellow.

Lexical items	Correspondence degree
<cause> *problems*	13.2
<there be> *problems*	12.4
<create> *problems*	11.8
problems <arise>	8.7
<give rise to> *problems*	8
<pose> *problems*	7.2
<present> *problems*	7
<raise> *problems*	4.7
<to be problematic>	4.5
<lead to> *problems*	4
<result in> *problems*	4

Lexical items	Correspondence degree
<zu> *Problemen* <führen>	12.1
Probleme <bringen>	10.4
Probleme <es gibt>	9.5
Probleme <verursachen>	8.4
Probleme <schaffen>	7.3
Probleme <bereiten>	7.3
Probleme <auftreten>	6.8
Probleme <entstehen>	6.6
Probleme <aufwerfen>	6.4
<problematisch sein>	5.7
Probleme <sich ergeben>	5.3
Probleme <darstellen>	4
<Ursache GEN|für> *Probleme*	3

Table 5: Correspondence degree values for English and German lexical items

The results indicate that the items with higher values of correspondence degree have a higher correspondence potential. In Section 5.3 we will see that the correspondence potential will play an important role in the creation of dictionaries.

5.2 Local grammars

In this section, local grammars of items that belong to the corresponding English and German domains {CAUSE PROBLEM} and {PROBLEM BEREITEN} will be discussed. These grammars will provide a description of textual contexts in which lexical items from these corresponding domains occur, and the description will reveal the functions of co-occurring items. This description is based on the observation of the distribution of English and German lexical items in the reference corpora ukWaC and deWaC.

At the most general level, three types of verbs can be distinguished in linguistic units from the two domains: transitive, intransitive and existential verbal expressions. As was seen above, the noun slot consists of the nouns <problem> or <difficulty> and they occur either in singular or plural. These nouns can be modified by adjectives or multi-word adjectival expressions, which can be classified into the following four groups according to their

function: INTENSIFIERS, QUANTIFIERS, SORTALS and COMPARATORS. Each group contains a restricted number of items.

INTENSIFIERS are lexical units that specify the degree of seriousness of problems. They can denote either a high or low degree of intensification. The lexical items from the category QUANTIFIERS indicate whether problems are large or small, and they occur mainly with the plural form of <problem> and <difficulty>. QUANTIFIERS are mostly one-word long and only seldom take the form of multi-word expressions. SORTALS specify the kinds of problems or difficulties encountered. The following groups of SORTALS occur in our data: health-related (e.g. <health>, <breathing>, <heart>, <skin>, <liver>, <eye>, <sleeping> or <dental>), communication-related (e.g. <language>, <communication>, <email>, <access> or <understanding>), security-related (e.g. <safety>, <security>, <health and safety>, <flooding>), technology-related (e.g. <technology>, <engineering>, <navigational>, <operational> or <technical>) and society and economy-related (e.g. <traffic>, <social>, <behavioural>, <unemployment>, <financial>, <environmental>, <economic>, <pollution> and <noise>). Finally, COMPARATORS denote whether problems or difficulties discussed in two different settings are of the same type. The most frequent types of INTENSIFIERS, QUANTIFIERS, SORTALS and COMPARATORS for English and German are displayed in Tables 6 and 7 respectively.

INTENSIFIERS	QUANTIFIERS	SORTALS	COMPARATORS
big	a few	access	additional
considerable	a great range of	behaviour	another
enormous	a lot of	communication	certain
great	a number of	engineering	different
huge	a series of	environmental	distinct
key	a small number of	ethical	further
large	all kind of	financial	new
major	all sort of	health	other
minor	fewer	legal	particular
serious	many	logistical	same
severe	more	management	similar
significant	numerous	noise	special
small	several	operational	typical
substantial	some	performance	unique
subtle		political	various
		pollution	
		practical	
		safety	
		security	
		technical	

Table 6: The most frequent types of modifiers of <problem> and <difficulty>

INTENSIFIERS	QUANTIFIERS	SORTALS	COMPARATORS
<arg>	<eine Reihe>	<beruflich>	<alt>
<echt>	<eine Vielzahl>	<finanziell>	<änlich>
<enorm>	<einig>	<gesellschaftlich>	<besonder>
<enorm>	<einzig>	<gesundheitlich>	<gewiß>
<erheblich>	<ein paar>	<Gesundheitsprobleme>	<gleich>
<ernst>	<kaum>	<intern>	<irgendwelch>
<ernsthaft>	<mehr>	<Kommunikationsproblem>	<neu>
<gering>	<viel>	<Kreislaufproblem>	<neuerlich>
<gewaltig>	<wenig>	<logistisch>	<speziell>
<gravierend>	<zahlreich>	<mental>	<unterschiedlich>
<groß>		<organisatorisch>	<weiter>
<Hauptproblem>		<politisch>	<zusätzlich
<Kernproblem>		<praktisch>	
<klein>		<psychisch>	
<massiv>		<rechtlich>	
		<sozial>	
		<Sprachproblem>	
		<strukturell>	
		<technisch>	
		<Verständnisproblem>	
		<wirtschaftlich>	

Table 7: The most frequent types of modifiers of <Problem> and <Schwierigkeit>

Verbal elements of the lexical items can also be modified by modal and adverbial expressions. The items from the local grammar classes PROBABILITY_OPERATORS, USUALITY and DURATION denote how likely it is that a problem will occur, how often it occurs and how persistent it is, respectively. Finally, the linguistic items from the class RECIPIENT indicate who will be affected by problems. A complete description of the local grammar of items from the current domain is displayed in Table 8. Square brackets denote optional and vertical bars alternative members.

Table 8: A local grammar of the lexical items from the CLDs {CAUSE PROBLEM} and {PROBLEM BEREITEN}

1. THING ^ [PROBABILITY_OPERATOR] ^ [CONTINUITY|USUALITY] ^ $CAUSE^{TR}$ ^ [RECIPIENT] ^ [INTENSIFIER|QUANTIFIER|SORTAL|COMPARATOR] ^ PROBLEM ^ [RECIPIENT]

2. $CAUSE^{EX}$ ^ [PROBABILITY_OPERATOR] ^ [CONTINUITY|USUALITY] ^ [INTENSIFIER|QUANTIFIER|SORTAL|COMPARATOR] ^ PROBLEM

3. [INTENSIFIER|QUANTIFIER|SORTAL|COMPARATOR] ^ PROBLEM ^ [PROBABILITY_OPERATOR] ^ $CAUSE^{INTR}$

Table 8: A local grammar of the lexical items from the CLDs {CAUSE PROBLEM} and {PROBLEM BEREITEN}

The above local grammar description summarises the textual context in which the lexical items from the domains occur. At the next stage, the distribution of lexical items is compared, with the aim of identifying individual differences and similarities between them. These differences and similarities are calculated in terms of association strength. I use the logDice test for this purpose. This test relies on the measurement of a coincidence index and measures the joint occurrence of *a* and *b* divided by the total occurrence of *a* and *b* separately in two samples (Dice, 1945; Rychlỳ, 2008). A full report of results is beyond the scope of the present paper, and I will discuss only a few examples of these differences and similarities.

It can be observed that the most typical adverbial modifiers of *problems* are the QUANTIFIERS <many>, <a lot of>, <a number of>, <several>, <some>, and <a few> which occur most often with <cause> and <lead to>. In some cases typicality depends on the word forms; <present> occurs with higher association strength with *problem* and *difficulty*, whereas <cause> is strongly associated with *problems* and *difficulties*. Table 9 displays the distribution of the RECIPIENT items and modal verbs in relation to the English lexical items. We can see that only half of the items colligate with the linguistic units from the class RECIPIENT and that the three most typical expressions that co-occur with modal verbs are <result in problem|difficulty>, <cause problem|difficulty> and <give rise to problem|difficulty>.

Lexical items	RECIPIENT	Co-occurrence with modal verbs
<result in problem|difficulty>		1
<cause problem|difficulty>	x	2
<lead to problem|difficulty>		3
<give rise to problem|difficulty>		4
<create problem|difficulty>	x	5
<problem|difficulty arise>		6
<present problem|difficulty>	x	7
<pose problem|difficulty>	x	8
<there be problem|difficulty>		9
<raise problem|difficulty>	x	10

Table 9: RECIPIENTS and modal verbs

To give two additional examples, <problem|difficulty arise> is the only item that occurs in the conditional expressions with <should> and <if>, and <create> collocates more typically than any other item with <more problem|difficulty than NP solve>.

Finally, two general tendencies can be observed in the current data. First, if A and B are two lexical items from the same domain and the former is more frequent than the latter, then the former item will co-occur with more than 50% of the collocates that occur with the latter one. The degree of overlap here increases as the difference between frequency values decreases. Second, if A and B are two lexical items from the same domain that share the same collocates and if A occurs with higher frequency, then this item will have a larger number of stronger collocates than B. The implication of these two tendencies is that the less frequent lexical items can usually be replaced by the more frequent items, provided they occur in the same textual context. The technical term used to describe this replaceability power of lexical items will be *substitution potential*. We will say that a lexical item has a higher substitution potential than other items if it can replace them in the same context and the resulting expression remains idiomatic.

5.3 Lexicographical relevance of the model

In this section I will discuss how the results from the interlingual and intralingual analysis can be used in the creation of onomasiological bilingual dictionaries.

Although the substitution and correspondence potential values tend to be related, they are not completely correlated. In other words, it does not necessarily follow that if a lexical item occurs in a greater range of contexts, it will correspond with a higher percentage to a larger number of lexical items from another language. We therefore need another model that will treat both variables equally.

I propose here an approach that relies on decision theory (e.g. Berger, 1985; Lehmann, 1950; Neumann and Morgenstern, 1953; Raiffa and Schlaifer, 1968). Decision-making is defined in decision theory in terms of weighing of risks against utilities. Risks represent types of uncertainty and insufficient information, which may lead to undesired outcomes. The reduction of risks is inversely proportional to the increase of utility values, because by increasing utilities we reduce risks. In decision theory, options are quantified and numerically represented; one weighs the available options and after having compared their values arrives at the option with the highest utility value.

Two kinds of risks are associated with the selection of synonymous expressions that correspond to lexical items from another language. First, lexical items do not occur with the same probability, and this means that a non-native speaker always bears the risk of selecting a non-typical option. Second, lexical items do not have the same correspondence potential, and

a language user risks choosing an option that might be at odds with the dominant understanding of cross-linguistic correspondence relations. In other words, when a user selects items from a lexical domain she risks either selecting an idiomatic expression that may not correspond suitably to the item from the source language, or selecting an item which may correspond but is not idiomatic. A model which suggests the most appropriate value must, therefore, take both variables into account. To enable this I propose a formula that adds up the values of the two variables and divides them by the number of variables (two). The option with the highest total sum will have the highest utility and lowest risk value.

I will provide one illustration. Let us imagine that a language user wants to translate <create> *problems* into German. In decision theory, alternatives and compared utilities are usually displayed in the form of decision tree diagrams, and one such diagram for our example can be seen in Figure 1.

Figure 1: A decision tree diagram for <create> *problems* and its German correspondences

The diagram indicates that there are five potential correspondences in German for the English lexical item. The German lexical units are represented in terms of the values of correspondence degree and collocation strength. It can be observed that *Probleme* <schaffen> has the highest correspondence potential and that <zu> *Problemen* <führen> has the highest substitution potential. In order to find the option that bears minimum risks, we sacrifice gains on each side and observe the final results (Table 10).

Translation correspondences of <create> problems	Utility values
Probleme <schaffen>	11.05
<zu> *Problemen* <führen>	10.9
Probleme <verursachen>	10.85
Probleme <bringen>	10.8
Probleme <entstehen>	10.6

Table 10: Utility values for German translation correspondences of <create> *problems*

The results indicate that the most appropriate translation of <create> *problems* is *Probleme* <schaffen>. Additional details about the use of the items can be added if the findings of the local grammar analysis are included. Thus, the verb <create> in this context occurs in passive constructions and colligates with RECIPIENTS. Similarly, the word form *problems* can be modified by QUANTIFIERS, INTENSIFIERS and the lexical item *no.* In addition, the whole expression occurs occasionally in the construction <create> *more problems than* <NP solve>. A comparison of the local grammars of English and German items suggests that <Schaffen> *Probleme* is the best option when <create> *problems* colligates with INTENSIFIERS, and when it participates in the construction <create> *more problems than* <NP solve>. On the other hand, <create> *no problems* is most appropriately translated as *keine Probleme* <verursachen>.

These results can be further used as resources for dictionary-makers of bilingual onomasiological dictionaries. The basic entries in such dictionaries would be corresponding CLDs in two or more languages. Lexical items in these domains would be ordered according to their correspondence and substitution potential. In addition, the local grammar information would specify how lexical items are typically used in the textual context. This information, along with the utility values of lexical items, would lift the dictionary user's burden and help in making "his or her own judgment on equivalences" (Atkins, 1996: 8). Thus, lexical domains would provide quick access to foreign terms corresponding to terms from one's

mother tongue, and the other way round. Additionally, CLDs might also serve as thesaurus-like entries.

An example of an entry of an electronic German-English onomasiological bilingual dictionary is provided in Figure 2. The English term discussed above (<create> problems) is displayed as a search term in the upper part of the Figure together with the name of the domain in which it occurs ({CAUSE PROBLEM}). The next section describes the local grammar of the search term and provides examples of how it is used. Finally, the second part of the entry contains the name of the corresponding German domain and the translation correspondences of the search term. The German translation correspondences are ordered according to their utility values, described above. In addition, by selecting the options 'local grammar' and 'example', a dictionary user has access to the local grammar and examples of the use of these German terms -. This entry serves only illustrative purposes, displaying in general terms how information obtained by means of the local grammar model can help dictionary users to identify corresponding lexical items in two languages, and to understand how these items are used in context.

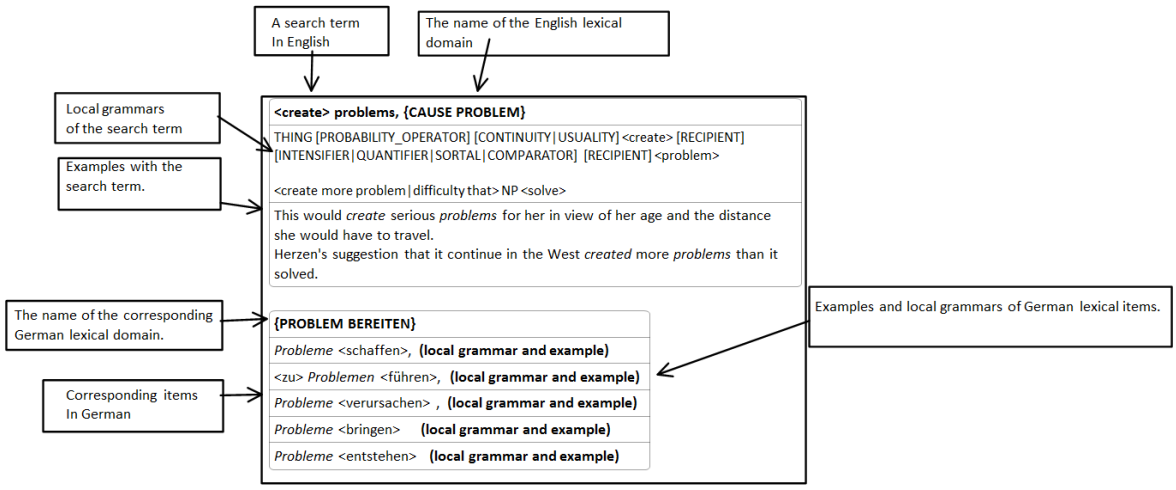

Figure 2: A sample dicionary entry for the search term <create> *problems.*

6. Conclusion

The aim of the present study was to develop an analytical framework that would enable compilation of a new type of onomasiological bilingual dictionary. The following two questions served as a point of departure:

i) How can we group semantically similar lexical items across languages?

ii) How can we distinguish between these lexical items?

The study conducted in Section 5 demonstrated that it was possible to provide positive answers to both questions, and to develop such a framework by combining the distributional approach with the language-in-use theory of meaning and corpus linguistic methodology. It was illustrated that corresponding lexical domains from two languages can be identified by applying a new cross-linguistic distributional hypothesis, and through the observation of the occurrence of lexical items in the parallel corpus. It was also demonstrated that lexical items from the same lexical domain can be differentiated in terms of their correspondence potential and local grammar features. Finally, the previous section showed that the results can be used in lexicography in the creation of bilingual onomasiological dictionaries. It can be concluded that the study shows that corresponding cross-linguistic semantic groups can be identified, and that synonymous lexical items can be differentiated purely on a distributional basis.

The present thesis was concerned only with the study of textual context, and I ignored the situational context. The model can, nevertheless, be extended to include this type of context as well. Such an analysis would explore how lexical items from the same lexical domain differ in terms of their occurrence in specific registers and genres. However, for such an analysis more research work should be done in the area of comparable studies of registers and genres, as no comparable cross-linguistic classifications of registers and genres are currently available.

Finally, analyses presented in the paper were conducted semi-automatically, and for large-scale studies the automatisation of the research process would be more than welcome, as it would speed up the analysis process and might potentially lead to further developments of the model. It would also make possible the creation of multilingual data bases that would serve as an alternative to existing platforms such as EuroWordNet or FrameNet. Unlike the former model, which might, as we saw in Section 2, multiply the number of senses for multiword expressions, the present model decreases their semantic ambiguity.

Acknowledgments

I am grateful to anonymous reviewers and the journal editor for their detailed and helpful comments. The usual disclaimer applies.

Dictionaries

Dornseiff, F. 2004. *Der deutsche wortschatz nach sachgruppen*. Berlin and Walter de Gruyter.

Kay, C., J. Roberts, M. Samuels and I. Wotherspoon (eds.) 2009. *Historical Thesaurus of the Oxford English Dictionary. With additional material from A Thesaurus of Old English.* Oxford: Oxford University Press.

McArthur, T. 1981. *Longman Lexicon of Contemporary English.* Harlow: Longman Harlow.

McCarthy, M. 1994 *Cambridge Word Routes Anglais-Français: Lexique thématique de l'anglais courant*. Cambridge: Cambridge University Press.

McCarthy, M. 1996 *Cambridge Word Routes Anglika-Ellinika*. Cambridge: Cambridge University Press.

Roget, P.M. 1852. *Roget's Thesaurus of English Words and Phrases.* London: Longman, Brown, Green, and Longmans.

Sanders, D. 1873. *Deutscher Sprachschatz geordnet nach Begriffen zur leichten Auffindung und Auswahl des passneden Ausdrucks*. Tübingen: Niemeyer.

Summers, D. 1993. *Longman Language Activator: The World's First Production Dictionary*. London: Harlow.

Walter, E. 1995 *Cambridge word selector: inglés-español: diccionario temático del inglés contemporáneo*. Cambridge: Cambridge University Press.

References

Apresjan, J. 2000. *Systematic Lexicography*. Oxford: Oxford University Press.

Atkins, B.T.S. 1994. 'Introduction.' In: Corréard, M.H. and V. Grundy (eds.) *Oxford-Hachette English-French dictionary, 1st edition*. Oxford: Oxford University Press and Hachette.

Atkins, B.T.S. 1996. 'Bilingual Dictionaries: Past, Present and Future.' In: Gellerstam, M., J. Jarborg, S-G. Malmgren, K. Noren, L. Rogstro and C.R. Papmehl (eds.) *Euralex'96 Proceedings I–II, Papers Submitted to the Seventh EURALEX International Congress on Lexicography in Goteborg, Sweden*. Gothenburg: Department of Swedish, Gothenburg University.

Atkins, B.T.S. and F.E. Knowles 1990. 'Interim Report on the EURALEX/AILA Research Project into Dictionary Use.' In *Proceedings of BudaLex*, 88, pp. 381–392.

Baker, G. and Hacker, P.M.S. 2009. *Wittgenstein: Rules, Grammar and Necessity: Volume 2 of an Analytical Commentary on the Philosophical Investigations, Essays and Exegesis §§ 185-242*. Oxford: Blackwell.

Barlow, M. 2008. 'Parallel Texts and Corpus-based Contrastive Analysis.' In: Gómez González, M., Mackenzie, L. and E. González Alvarez (eds.) *Current Trend in Contrastive Linguistics.* Amsterdam: John Benjamins, pp. 101-121.

Baxter, J. 1980. 'The Dictionary and Vocabulary Behavior: A Single Word or a Handful?' *Tesol Quarterly*, 14, pp. 325–336.

Bendix, E.H. 1971. 'The Data of Semantic Description.' In: Steinberg, D.S. and Jakobovits, L.A. (eds.) *Semantics: An interdisciplinary Reader in Philosophy, Linguistics, and Psychology.* Cambridge: Cambridge: University Press, pp. 393–409.

Berger, J.O. 1985. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer.

Boas, H.C. 2002. 'Bilingual FrameNet Dictionaries for Machine Translation.' In: González-Rodríguez, M. and C.P. Suárez Araujo (eds.) *Proceedings of the Third International Conference on Language Resources and Evaluation*. Las Palmas, Spain, IV, pp. 1364-1371.

Boas, H.C. 2005. 'Semantic Frames as Interlingual Representations for Multilingual Lexical Databases.' *International Journal of Lexicography*, 18 (4), pp. 445–478.

Braasch, A. 1994. 'There's no Accounting for Taste - Except in Dictionaries.' *Proceedings. Amsterdam: EURALEX*. Amsterdam, pp. 45-55.

Church K.W., Gale, W., Hanks, P., Hindle, D. and R. Moon 1994. 'Lexical Substitutability.' In: Atkins B.T.S. and A. Zampolli *Computational Approaches to the Lexicon.* Oxford: Clarendon Press, pp. 153-177.

Claes, F. 1977. *Bibliographisches Verzeichnis der deutschen Vokabulare und Wörterbücher, gedruckt bis 1600*. Hindelsheim: Georg Olms Verlag.

Cruse, A. D. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

Desgraupes, B. and S. Loiseau 2012. *Introduction to the rcqp package*. Available from: <http://cran.r-project.org/web/packages/rcqp/vignettes/rcqp.pdf> [Accessed 25 July 2015]

Dice, L.R. 1945. 'Measures of the Amount of Ecologic Association Between Species.' *Ecology*, 36(3), pp. 297–302.

Dixon, R.M.W. 1971. 'A Method of Semantic Description.' In: Steinberg, D. and L. Jakobovits, (eds.) *Semantics*. Cambridge: Cambridge University, pp. 436–471.

Durrell, M. 1981. 'Contrasting the Lexis of English and German.' In: Russ, C.V.J. (ed.) *Contrastive Aspects of English and German.* Heidelberg: Groos, pp. 35-54.

Durrell, M. 2000. *Using German Synonyms*. Cambridge: Cambridge University Press.

Dyvik, H. 1998. 'A Translational Basis for Semantics.' *Language and Computers*, 24, pp. 51–86.

Dyvik, H. 2004. 'Translations as Semantic Mirrors: from Parallel Corpus to WorldNet.' In: Aijmer, K. and B. Altenberg (eds.) *Language and Computers, Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23. Göteborg 22-26 May 2002.* Amsterdam: Rodopi, pp. 311-326.

Dyvik, H. 2005. 'Translations as a Semantic Knowledge Source.' *Proceedings of the second Baltic Conference on Human Language Technologies.* Tallin, Estland, pp. 27–38.

Evans, V. and Green, M. 2006. *Cognitive Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

Evert, S. and A. Hardie 2011. 'Twenty-first Century Corpus Workbench: Updating a Query Architecture for the New Millennium.' *Corpus Linguistics Conference*. Birmingham, University of Birmgingham.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT. Press.

Ferraresi, A., E. Zanchetta, M. Baroni and S. Bernardini 2008. 'Introducing and Evaluating UKWAC, a Very Large Web-derived Corpus of English.' *Proceedings of the 4th Web as Corpus Workshop (WAC-4. Can We Beat Google)*. Marrakech, Morocco, pp. 47-54.

Fillmore, C.J. and B.T.S. Atkins 1992. 'Toward a Frame-based Lexicon: The Semantics of RISK and its neighbors' In: Lehrer, A. and E.F. Kittay (eds.) *Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization.* London: Routledge, pp. 75-102.

Firth, J.R. 1968. 'A synopsis of Linguistic Theory 1930-1955.' In: Palmer, F.R. (ed.) *Selected Papers of J.R. Firth 1952-1959*. London: Longman, pp.168-205.

Geeraerts, D. 2010. *Theories of Lexical Semantics.* Oxford and New York: Oxford University Press.

Goddard, C. and N. Thieberger 1997. 'Lexicographic Research on Australian Aboriginal Languages, 1969-1993' In: Tryon, D. and M. Walsh (eds.) *Boundary Rider: Essays in Honour of Geoffrey O'Grady*. Canberra: Pacific Linguistics, Research School of Pacific and Asian Studies, Australian National University, pp. 175-208.

Gordon, W.T. 2003. 'Semantic Theories in 20th-century America: An Overview of Approaches Outside Generative Grammar' In: Wiegand, H.E. (ed.) *History of the Language Sciences.* Berlin: de Gruyter, pp. 2213-2229.

Grondelaers, S. and D. Geeraerts 2003. 'Towards a pragmatic model of cognitive onomasiology.' In: Cuyckens, H., Dirven, R. and J. Taylor (eds.). *Cognitive Approaches to Lexical Semantics*. Berlin: Mouton de Gruyter, pp 67-92.

Gross, M. 1993. 'Local Grammars and their Representation by Finite Automata.' In: Hoey, M. (ed.) *Data, Description, Discourse. Papers on the English Language in Honour of John McH Sinclair*. London: Collins, pp. 26–38.

Grosz, B.J. 1982. 'Discourse Analysis.' In: Kittredge, R. and J. Lehrberger (eds.) *Sublanguage: Studies of Language in Restricted Semantic Domains*. Berlin and New York: Mouton de Gruyter, pp. 138–174.

Hanks, P. 2004. 'The Syntagmatics of Metaphor and Idiom.' *International Journal of Lexicography*, 17(3), pp. 245–274.

Hanks, P. and Pustejovsky, J. 2005. 'A Pattern Dictionary for Natural Language Processing.' *Revue Française de linguistique appliquée*, 10(2), pp. 63–82.

Harris, R. 2005. *The Semantics of Science*. London: Continuum International Publishing Group.

Harris, Z.S. 1952. 'Discourse Analysis.' *Language*, 281., pp. 1–30.

Harris, Z.S. 1954. 'Distributional Structure.' *Word*, 102/3., pp. 146-162.

Harris, Z. S. 1970. *Papers in Structural and Transformational Linguistics*. Dordrecht: Reidel.

Harris, Z. S. 1988. *Language and information*. New York: Columbia University Press.

Harris, Z.S., M. Gottfried, T. Ryckman, Jr. P. Mattick, A. Daladier, T.N. Harris and S. Harris 1989. *The Form of Information in Science: Analysis of an Immunology Sublanguage.*

Boston Studies in the Philosophy of Science, 104. Dordrecht and Boston: Kluwer Academic Publishers.

Hartmann, R.R.K. and James, G. 1998. *Dictionary of Lexicography*. London and New York: Routledge.

Heid, U. 1996. 'Creating a Multilingual Data Collection for Bilingual Lexicography from Parallel Monolingual Lexicons.' *Euralex'96 Proceedings*, pp. 573-590

Hüllen, W. 1999. *English Dictionaries, 800-1700: The Topical Tradition*. Oxford: Oxford University Press.

Hüllen, W. 2009. *Networks and knowledge in Roget's thesaurus*. Oxford: Oxford University Press.

Hunston, S. and J.M. Sinclair 2000. 'A Local Grammar of Evaluation.' In: Hunston S. and G. Thomson (eds.) *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: Oxford University Press, pp. 74–101.

Kilgarriff, A. and I. Kosem 2012. 'Corpus Tools for Lexicographers.' In: Granger, S. and M. Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford University Press, pp. 31-57

Kittredge, R. 1982. 'Variation and Homogeneity of Sublanguages.' In: Kittredge, R. and J. Lehrberger (eds.) *Sublanguage: Studies of Language in Restricted Semantic Domains*. Berlin and New York: Mouton de Gruyter, pp. 107–137.

Koehn, P. 2005. 'A Parallel Corpus for Statistical Machine Translation.' *Proceedings of MT Summit* **X**. Phuket, Thailand, pp. 79-86.

Lehmann, E.L. 1950. 'Some Principles of the Theory of testing hypotheses.' *The Annals of Mathematical Statistics*, 211, pp. 1–26.

Lehrer, A. 1974. *Semantic fields and Lexical*. Amsterdam: North-Holland.

Lyons, J. 1977. *Semantics (Vols I & II)* Cambridge**:** Cambridge University Press.

Lyons, J. 1995. *Linguistic Semantics: An Introduction*. Cambridge: Cambridge University Press.

Martin, S.E. 1967. 'Selection and Presentation of Ready Equivalents in a Translation Dictionary.' In: Householder, F.W.S. and S. Saporta (eds.) *Problems in Lexicography*. Bloomington: Indiana University, pp.153-159.

McArthur, T. 1986. *Worlds of reference: Lexicography, Learning and Language from the clay tablet to the computer*. Cambridge: Cambridge University Press.

McArthur, T. 1998 *Living Words: Language, Lexicography and the Knowledge Revolution*. Exeter: University of Exeter Press.

Murphy, M. L. 2003. *Semantic relations and the lexicon: Antonymy, synonymy, and other paradigms*. Cambridge, UK: Cambridge University Press

Nord, B. 2002. *Hilfsmittel beim Übersetzen: Eine empirische Studie zum Rechercheverhalten professioneller Übersetzer.* Frankfurt am Main: Lang.

Raiffa, H. and R. Schlaifer 1968. *Applied Statistical Decision Theory*. Boston: Harvard University.

Rychlỳ, P. 2008. 'A Lexicographer-friendly Association Score.' *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN, Brno, Czech Republlic*, Masaryk University, pp. 6–9.

Salkie, R. 1995. 'INTERSECT: A Parallel Corpus Project at Brighton University.' *Computers and Texts*, 9, pp. 4–5.

Siepmann, D. 2005. *'Collocation, Colligation and Encoding Dictionaries. Part I: Lexicological Aspects.' International Journal of Lexicography* 18.4. 409–443.

Siepmann, D. 2006. 'Collocation, colligation and encoding dictionaries. Part II: Lexicographical aspects.' *International Journal of Lexicography*, 19 1. pp. 1— 39.

Sinclair, J.M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J.M. 1996a. 'An International Project in Multilingual Lexicography.' *International Journal of Lexicography*, 9 (3), pp. 179–196.

Sinclair, J.M. 1996b. 'Corpus to Corpus: A Study of Translation Equivalence.' *International Journal of Lexicography*, 9 (3), 171–178.

Sinclair, J.M. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Snell-Hornby, M. 1984. 'The Bilingual Dictionary: Help or Hindrance.' *LEXeter'83 Proceedings*. Tubingen: Max Niemeyer Verlag, pp. 274–281.

Snell-Hornby, M. 1987. 'Towards a learners' Bilingual Dictionary.' In: Cowie, A.P. (ed.) *The dictionary and the Language learner*. Tubingen: Max Niemeyer Verlag, pp. 159-170.

Snell-Hornby, M. 1990. 'Dynamics in Meaning as a Problem for Bilingual Lexicography.' In: Tomaszczyk, J. (ed.) *Meaning and Lexicography*. Amsterdam: John Benjamins, pp. 209-225.

Sterkenburg, P. van 2003. 'Onomasiological specifications and a concise history of onomasiological dictionaries.' In: Sterkenburg, P. van (ed.) *A Practical Guide to Lexicography*. Amsterdam: John Benjamins, pp. 127– 143.

Teubert, W. 2004. 'Language and Corpus Linguistics.' In: Halliday, M.A.K., Teubert, W., Yallop, C. and A. Čermáková (eds.) *Lexicology and Corpus Linguistics*. London and New York: Continuum, pp. 73–112.

Tomaszczyk, J. 1979. 'Dictionaries: Users and Uses.' *Glottodidactica*, 12, pp. 103–119.

Trier, J. 1931. *Der deutsche Wortschatz im Sinnbezirk des Verstandes*. Heidelberg: C. Winter.

Trklja, A. 2013. *A Corpus linguistics study of translation correspondences in English and German.* PhD Thesis, University of Birmingham.

Van Roey, J. 1990. *French-English Contrastive Lexicology: An Introduction*. Leuven: Peeters Publishers.

Viberg, Å. 1983. 'The Verbs of Perception: A Typological Study.' *Linguistics* 211, pp. 123-62.

Viberg, Å. 1993. 'Cross-Linguistic perspectives on Lexical Organization and Lexical Progression.' In Hyltenstam, K. (ed.) *Progression and Regression in Language: Sociocultural, Neuropsychological, and Linguistic Perspectives*. Cambridge: Cambridge University Press. pp. 340-383.

Von Neumann, J. and O. Morgenstern 1953. *Theory of Games and Economic Behavior.* Princeton, NJ: Princeton University Press.

Vossen, P. 2004. 'EuroWordNet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual-index.' *International Journal of Lexicography*. 2004. 17. (2), pp.161-173.

Waismann, F. 1965. *The Principles of Linguistic Philosophy*. London: Macmillan.

Wierzbicka, A. 1987. *English Speech Act verbs: A Semantic Dictionary*. Sydney: Academic Press.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Basil Blackwell.

Zgusta, L. 2006. *Lexicography Then and Now: Selected Essays*. Lexicographica. Series maior, 129. Tübingen: Max Niemeyer.

Corpora used in the present study

British National Corpus
DeWaC German Web Corpus
English-German/German-English Europarl Parallel Corpus
UKWaC British English web Corpus
OpenSubtitles corpus

Appendix

Table I: English lexical items from the CLD {CAUSE PROBLEM} and the corresponding German lexical items.

English as the source language		
<create problem|difficulty> 456: <Problem|Schwierigkeit schaffen> 107, <zu Problem|Schwierigkeit führen> 58, <Problem|Schwierigkeit bringen> 46, <Problem|Schwierigkeit verursachen> 41, <Problem|Schwierigkeit bereiten> 23, <Problem|Schwierigkeit entstehen> 16, <Problem|Schwierigkeit aufwerfen> 13, <Problem|Schwierigkeit auftreten 7, <problematisch sein> 7, <Problem|Schwierigkeit darstellen> 5, <es gibt <Problem|Schwierigkeit> 4, <vor Problem|Schwierigkeit stellen> 3, <Problem zur Folge haben> 3.	**<cause Problem|Schwierigkeit 569:** <Problem|Schwierigkeit verursachen> 110, <zu Problem|Schwierigkeiten führen> 79, <Problem|Schwierigkeit bereiten> 62, <Problem|Schwierigkeit bringen 61, <Problem|Schwierigkeit aufwerfen> 27, <problematisch sein>24, <Probleme darstellen> 16, <Problem|Schwierigkeit schaffen> 15, <Problem|Schwierigkeit hervorrufen> 14, <Problem|Schwierigkeit ergeben> 14, <Problem|Schwierigkeit entstehen> 10, <vor Problem|Schwierigkeit stellen> 9, <Problem|Schwierigkeit auftreten> 8, <es gibt Problem|Schwierigkeit> 7, <Ursache für die Probleme> 3, <Ursache der Probleme> 3.	**<pose problem|difficulty> 220:** <Problem|Schwierigkeit aufwerfen> 32, <Problem|Schwierigkeit darstellen> 37, <Problem|Schwierigkeit bereiten> 22, <problematisch sein> 18, <Problem|Schwierigkeit bringen 15, <vor Problem|Schwierigkeit stellen> 12, <zu Problem|Schwierigkeit führen> 12, <vor Problem|Schwierigkeit stellen> 8, <Problem| Schwierigkeit schaffen> 7, <es gibt Problem|Schwierigkeit 7>, <Problem|Schwierigkeit ergeben> 5, <Problem|Schwierigkeit stellen sich> 4, <Problem|Schwierigkeit verursachen> 3, <Problem|Schwierigkeit auftreten> 2, <stellt sich Problem> 2.
<present problem|difficulty> 125: <Problem|Schwierigkeit darstellen> 20, <vor Problem|Schwierigkeit stellen> 12, <Problem|Schwierigkeit bereiten> 12, <Problem|Schwierigkeit bringen> 12, <Problem|Schwierigkeit aufwerfen> 11, <Problem|Schwierigkeit verursachen> 8, <Problem|Schwierigkeit sich ergeben> 6, <problematisch sein> 5, <es gibt Problem|Schwierigkeit> 5, <mit Problem verbunden sein> 2, <Problem|Schwierigkeit aufweisen> 2, <stellt sich Problem> 3	**<problem| difficulty arise>688:** <Problem|Schwierigkeit entstehen> 92, <Problem|Schwierigkeit auftreten> 90, <Problem|Schwierigkeit sich ergeben> 63, <Problem|Schwierigkeit sich stellen> 46, <es gibt Problem|Schwierigkeit> 27, <Problem|Schwierigkeit auftreten> 12, <Problem|Schwierigkeit schaffen> 10, <Problem|Schwierigkeit bringen> 8, <problematisch sein> 9 , <Problem|Schwierigkeit darstellen 5, <zu Problem|Schwierigkeit führen> 3, <Problem|Schwierigkeit	**<give rise to problem|difficulty> 79:** <zu Problem|Schwierigkeiten führen> 17, <Problem|Schwierigkeit entstehen> 10, <Problem|Schwierigkeiten aufwerfen> 6, <mit <Problem|Schwierigkeiten verbunden sein> 5, <es gibt Problem|Schwierigkeit> 5, <Problem|Schwierigkeit 4, <Problem|Schwierigkeit schaffen> 3, <Problem|Schwierigkeit verursachen> 3, <Probleme|Schwierigkeit sich ergeben>3 , <Problem|Schwierigkeiten bereiten> 2,

Table II: German lexical items from the CLD {PROBLEM BEREITEN} and the corresponding English lexical items.

German as the source language			
<zu Problem|Schwierigkeit führen> 322: <lead to problem|difficulty> 86, <cause problem|difficulty> 60, <create problem|difficulty> 30, <give rise to problem|difficulty> 13, <pose problem|difficulty> 12, <raise problem|difficulty> 9, <there be problem|difficulty> 9, <result in problem|difficulty> 7, <present problem|difficulty> 5, <to be problematic> 5.	**<Problem|Schwierigkeit schaffen> 212:** <create problem|difficulty> 128, <cause problem|difficulty> 25 <raise problem|difficulty> 10 , <lead to problem|difficulty>7, <there be problem> 6, <give rise to problem|difficulty> 5, pose problem|difficulty> 4.	**<Problem|Schwierigkeit verursachen> 271:** <cause problem|difficulty> 135, <create problem|difficulty> 51, <present problem|difficulty> 10, <give rise to problem|difficulty> 8, <raise problem|difficulty> 7, <there be problem 4 <pose problem> 4, <lead to problem|difficulty> 3.	**<Problem|Schwierigkeit (mit sich) bringen> 196:** <cause problem|difficulty> 42, <create problem|difficulty> 34, <bring problem|difficulty> 18, <raise problem|difficulty> 15, <pose problem|difficulty> 11, <present problem|difficulty> 7, <there be problem|difficulty> 7, <lead to problem|difficulty> 6, <problem|difficulty arise> 4.
<Problem|Schwierigkeit> bereiten> 312: <cause problem|difficulty> 86, <create problem|difficulty> 39, <to be difficult> 29, <pose problem|difficulty> 21, <there be problem|difficulty> 14, <have problem|difficulty>12, <present problem|difficulty> 9, <result in problem|difficulty> 8, <raise problem|difficulty> 7, <give rise to problem|difficulty> 3, <problem arise> 3.	**< Problem|Schwierigkeit auftreten> 112:** <problem|difficulty arise> 33, <there be problem|difficulty> 22, <problem|difficulty occur> 8, <create problem|difficulty> 5, <cause problem|difficulty> 4, <present problem|difficulty> 3.	**<Problem|Schwierigkeit sich ergeben> 101:** <problem|difficulty> arise>22, <there be problem|difficulty> 14, <cause problem|difficulty> 13 <raise problem|difficulty> 4, <present problem|difficulty> 4, <pose problem|difficulty> 4, <create problem|difficulty> 4, <give rise to problem|difficulty> 4, <have problem|difficulty> 3.	**<Problem|Schwierigkeit entstehen> 133:** <cause problem|difficulty> arise>51, <there be problem|difficulty> 26, <create problem|difficulty> 14, <cause problem|difficulty> 9, <give rise to problem|difficulty> 9, <lead to problem|difficulty> 3
<Problem|Schwierigkeit aufwerfen> 215: <raise problem|difficulty> 76, <pose problem|difficulty 30, <cause problem|difficulty> 27, <create	**<problematisch sein> 297:** <to be problematic> 138, <there be problem|difficulty> 45, <cause problem|difficulty> 35,	**<Problem darstellen> 397:** <to be a problem> 176, <to be an issue> 63, <present problem> 24, <pose problem> 17, <cause problem> 17,	**<es gibt Problem|Schwierigkeit> 1387** <there be problem|difficulty> 738, <have problem> 39,<cause